R E P O R T R E S U M E S

ED 010 978 RE 000 015

ENGLISH WORDS OF VERY HIGH FREQUENCY.

BY- CARD, WILLIAM MCDAVID, VIRGINIA

PUB DATE MAY 66

EDRS MF-$0.09 HC-$0.44 11P.

DESCRIPTORS- *READING RESEARCH, *STRUCTURAL ANALYSIS, *COMPARATIVE ANALYSIS, *SYNTAX, *LANGUAGE, ENGLISH, VOCABULARY, ADULT, READABILITY

THE BIAS OF THE FREQUENCY OF THE 122 MOST COMMONLY USED ENGLISH WORDS WAS STUDIED. THE METHOD USED TO ASSEMBLE THESE DATA IS DESCRIBED FULLY. THE MOST FREQUENTLY USED WORDS WERE TAKEN FROM A DISSERTATION BY GEORGE K. MONROE, "PHONEMIC TRANSCRIPTION OF GRAPHIC POSTBASE AFFIXES IN ENGLISH," GODFREY DEWEY, "RELATIVE FREQUENCY OF ENGLISH SPEECH SOUNDS," MILES L. HANLEY, "WORD INDEX TO JAMES JOYCE'S ULYSSES," AND HENRY D. RINSLAND, "A BASIC VOCABULARY OF ELEMENTARY SCHOOL CHILDREN." ALL BUT THE RINSLAND LIST WERE TAKEN FROM ADULT READING MATERIAL. WORDS WERE ARRANGED IN RANK ORDER. THE FREQUENCIES OF THE VARIOUS STUDIES WERE RECORDED. DATA WERE COMPARED ACCORDING TO SPECIFIC WORDS AS WELL AS ACCORDING TO STRUCTURE WORDS. REFERENCES ARE INCLUDED. THIS ARTICLE IS PUBLISHED IN "COLLEGE ENGLISH," MAY 1966. (BK)

College English, May 1966, 27, pp. 596-604

ED010978

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

RE 000 015

English Words of Very High Frequency

WILLIAM CARD AND VIRGINIA MCDAVID

I

A COMPARISON OF the two lists of words of highest frequency drawn from two sizable corpora of English can reveal some of the peculiarities or biases of the corpora. In making such comparisons,[1] we have discovered that the bias is much more apparent if three or four such lists are compared with each other. The bias can be detected by discrepancies in the frequencies of some words or by comparing the lists of the hundred (or more) words of highest frequency.

This paper will illustrate the method and will demonstrate that the bias of a corpus of 100,000 words or more reaches even into the 122 words of highest frequency. In view of the innumerable man-hours that have gone into piling up frequency counts from corpora numbering several millions of words in the effort to establish a bias-free frequency list, this is an important finding. It is made possible in part by the publication of the 122 words of highest frequency in the first 285,062 collected at Brown University for the Standard Corpus of Edited American English, of which the components were first put in print in 1961.

The first 285,062 words put on computer tape were the material of a doctoral dissertation by George K. Monroe.[2] We will call this portion of the Standard Corpus the Monroe corpus or M. For the purposes of the dissertation it was convenient to have the computer make a frequency count. The first page of the computer printout held 122 words and their frequencies. They are displayed in Table V[3] of the description of the Standard Corpus given by W. Nelson Francis in *College English* 26 (January 1965), 267-73. In Table I Francis lists the categories of publications sampled for the whole Standard Corpus and the number of 2000-word excerpts (to a total of 500) taken in each main category. We have not been able to learn from the Department of Linguistics at Brown which excerpts were included in the Monroe corpus and will be unable to account for the bias of M which our comparisons reveal.

In Table I we array the 122 words in a column; in column M we give the rank order of each and in column f the frequency. In three columns alongside M we give the rank order of the same words in three other frequency counts.

Column D is derived from Godfrey Dewey, *Relativ[e] Frequency of English Speech Sounds*, Harvard University Press, Cambridge, 1923. Like M, Dewey's corpus was a composite. It was assembled in the spring of 1918 and was chiefly but

[1] For another example see our "Frequencies of Structure Words in the Writing of Children and Adults," *Elementary English* 42 (December 1965), 878-882 and 894.

Mr. Card and Mrs. McDavid are professors of English at Illinois Teachers College: Chicago--South, formerly Chicago Teachers College South. *A first version of this paper was read at the First Regional Meeting of the Chicago Linguistic Society, April 4, 1965; a revised version was read at the Midwest Modern Language Association convention in Chicago May 7, 1965.*

[2] *Phonemic Transcription of Graphic Posthase Affixes in English: A Computer Problem.* Brown University, June 1965.

[3] By mistake the word at order 92 is given as *his* instead of *Mrs.* We have also corrected (from Monroe's thesis) the mistaken estimate of the size of M. A detailed description of the Standard Corpus is given in the *Manual of Information* which can be obtained from the Department of Linguistics at Brown.

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY Editor, COLLEGE ENGLISH

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

not strictly contemporary, some 10 per cent of the 107,138 running words having been composed before the twentieth century. Newspaper excerpts supplied 30 per cent of the corpus, magazines 25 per cent. Only 10 per cent was from fiction and 5 per cent from drama. Dewey lists the thousand words of highest frequency in his corpus. In column D we give the rank order in Dewey's list of the 122 cited from M.[4]

Column U shows the rank order of the same words as given in Miles L. Hanley, *Word Index to James Joyce's Ulysses*, Madison, Wisconsin, 1937 [mimeographed],[5] which notes the frequency of the 260,430 words, name initials, etc. in *Ulysses*, which Joyce composed in the years 1914-21.

Column R furnishes for the same words the rank order derived from Henry D. Rinsland, *A Basic Vocabulary of Elementary School Children*, Macmillan, New York, 1945, which lists alphabetically with their frequencies the 14,571 words occurring three or more times in any one grade from the first to the eighth in a corpus of 6,012,359 running words of schoolchildren's writing (including 4630 pages of recorded conversation of children in first grade) collected in the spring of 1937 from schools in all parts of the country.

For the reader's ease we carry out the exposition in this paper in terms of rank order rather than raw frequency: the numbers are smaller and easier to follow. In making a statistical check of the judgments in section II, we took the raw frequencies, adjusted them to equate with a corpus of 285,000 words, and ran a chi-square test. Except for those otherwise labeled, all the statements of significant differences in rank orders in the next section are valid for the raw frequencies well within the .001 level of confidence, which is to say that such differences in frequency (and hence rank order) would not turn up as often as once in a thousand times by mere chance. The excepted data are labeled for the .01, .02, or .05 levels of confidence, which relate respectively to a chance occurrence once in a hundred, once in fifty, or once in twenty times.[6]

TABLE 1

R	U	D	M		f
1	1	1	1	the	20,172
10	2	2	2	of	10,427
3	3	3	3	and	7,625
4	5	4	4	to	7,565
5	4	5	5	a	6,322
8	6	6	6	in	6,160
11	19	9	7	is	3,417
20	9	7	8	that	2,962
17	15	11	9	for	2,941
9	12	8	10	it	2,213
28	11	16	11	with	2,040
62	24	14	12	as	2,011
12	13	13	13	was	1,967
18	14	18	14	on	1,953
56	34	12	15	be	1,928
16	7	17	16	he	1,682
95	22	20	17	by	1,658
43	64	23	18	this	1,590
15	54	24	19	are	1,533
38	21	22	20	at	1,485
63	8	26	21	his	1,395
59	26	33	22	from	1,351
13	44	19	23	have	1,259
101	28	30	24	or	1,216
37	32	21	25	not	1,183
42	40	27	26	but	1,103
134	53	31	27	which	1,067
99	46	38	28	an	1,030
24	78	32	29	will	939
77	91	35	30	has	938
31	43	36	31	one	930
19	27	28	32	they	875
6	16	15	33	you	865
21	36	34	34	had	805

[4] Dewey dropped abbreviations from his list. We have inserted *Mr.* and *Mrs.* at their proper places and adjusted the rank orders under D to include them.

[5] In Appendix II of this volume Martin Joos gives the rank order in U of the first 100 words in Dewey's list. This suggested to us that it would be interesting to add M to the comparison. A former colleague, Dr. Alda Raulin, suggested that we add the Rinsland frequencies.

[6] Our thanks are due to Professor Leonard Newmark for pointing out the desirability of making a statistical check of our findings and to our colleague Professor Carl Clark for helping us to choose a test appropriate to our data.

TABLE 1 *(cont'd)*

R	U	D	M		f
7	69	25	35	we	803
30	20	29	36	all	762
53	55	43	37	were	723
92	41	41	38	their	716
51	73	49	39	would	711
141	60*	50	40	who	697
136	86*	58*	41	more	675
68	128	65	42	can	662
112	126*	39	43	been	627
148	185*	120	44	new	625
35	42	42	45	there	613
80	51	46	46	if	602
25	50	51	47	when	597
312	110*	60	48	its	576
228	165*	66	49	than	546
163	47	40	50	no	545
72	23	75	51	said	517
32	48	44	52	so	504
195	102*	67	53	only	491
104	85*	70	54	other	490
272	344	76	55	may	480
45	83*	80*	56	some	475
71	31	48	57	what	468
247	270*	81*	58	these	455
73	18	52	59	him	439
40	33	64*	60	out	435
60	17	54	61	her	433
58	52	78*	62	about	432
114	79	71	63*	into	431
48	37	62	64*	up	431
52	95	37	65	our	425
113	98	89	66	first	422
86	72	83	67	two	413
33	59*	56*	68	your	405
49	74	61	69	time	404
202	182	109*	70	most	400
61	45	53	71	them	399
57	66	63*	72	do	389
22	25	68	73	she	388
90	57	79*	74	over	387
185	292*	152	75	also	378
410	208*	88	76	such	377
149	138*	57*	77	any	372
105	323*	110	78	many	364
46	49	97	79	then	343
100	88	87	80	could	337
82	68	105	81	after	325
76	63	59	82	now	324
130	148*	138*	83	last	318
176	207*	124	84	years	311
97	145	69	85	made	308
504	214*	122*	86	even	302
231	120	73*	87	must	285
400	183*	115*	88	world	279
70	84*	107	89	good	278
116	70	77	90	man	277
188	313*	96	91	should	275
309*	126*	236*	92	Mrs.	269
301	39	82	93	Mr.	268
491	87	112*	94	those	263

TABLE 1 *(cont'd)*

R	U	D	M		f
177	108	135*	95*	through	259
248	281	159*	96*	each	259
89	121*	102*	97*	because	259
14	35	45	98	my	258
171	360*	174*	99	year	255
568*	1403*	282*	100	state	254
158	121*	85	101	before	253
115	354	74*	102*	people	252
39	38	98	103*	like	252
55	99	92	104	how	249
119	162*	111*	105	much	245
154	105*	113*	106	way	243
132	89*	132*	107*	where	241
111	157	108*	108*	make	241
96	107	130*	109	just	237
93	96	99	110	well	236
23	117*	84	111	very	232
293	117*	104*	112	under	219
69	105*	117*	113	day	218
180	393*	133*	114	work	216
322	682*	283*	115	use	213
144	167*	144*	116	three	211
103	65	184*	117	too	207
461	155*	103*	118	being	203
535*	180*	143*	119	own	202
469	493*	199*	120	since	200
309*	137	180*	121	still	199
284	331*	231*	122	used	198

*Approximate: two or more words of the same frequency

II

The most common words are the ones whose rank order in a frequency list are most firmly fixed for a corpus of a given kind. Yet bias may show itself in the first few ranks if the corpus differs sufficiently from the norm. We note that M and D agree on the ranking of the first six words; U agrees to the list but reverses the rank of *a* and *to;* but R agrees to only four of the list, *in* and especially *of* dropping down to positions lower in order. With frequencies of the magnitude of those occurring in the first six ranks, the chance that the shifts in U and R are a random effect is a great deal less than one in a thousand.

M and D agree on 35 of the first 36 words, as indicated by the fact that only one number of the first 36 under D is higher than 36. The word not agreed on is *I*, which stands at order 10 in D and U and at order 2 in R but is not included in

the 122 of M. Under U, 10 of the first 36 numbers are higher than 36, and under R, 12. Further study confirms the first impression that D is closer to M than the other two lists are, that R is farthest from M, and that U is between R and D but somewhat closer to R.

The farther down the list one goes, the greater the discrepancies in the rank orders. The point of greatest diversion happens to be at the word *state*, order 100 in M and 282 in D. The discrepancy of 182 rank orders is greater (as between M and D) than for any other word on the list and approached only by *use* at rank 115. The rank order of *state* in R and U is also very much lower than in M. The second most extreme discrepancy between U and M is also for the word *use*. Evidently the frequencies of *state* and *use* are abnormally high in M.

Other words of markedly higher frequency in M than in the other lists are *new* 44, *years* 84, *year* 99, *used* 122, *also* 75, *even* 86, *last* 83, *each* 96, and *Mrs.* 92. For *Mrs.* the Dewey frequency is undoubtedly rather low: only the 15 per cent of fiction and drama in his corpus would be likely to heighten the frequency of the title. The Standard Corpus includes 5 excerpts from women's magazines and 3 from newspaper society pages; the fiction comprises 25 per cent of the whole corpus. A disproportionate amount of this matter must have got into the Monroe corpus. We predict that the frequency list of the Standard Corpus will place *Mrs.* below *Mr.* but not as low as the Dewey order of 236. The other words listed in this paragraph will also probably have a lower relative frequency in the count of the Standard Corpus.

Not so widely divergent from the other orders but still of notably higher order in M are the determiners *some* 56 and (at the .02 level) *these* 58; the numerative adjectives *more* 41, *first* 66, *most* 70, and (at the .01 level) *two* 67 and (at .02) *other* 54; the modals *can* 42 and *may* 55 (but only at .01). The words of notably low frequency in M are *your* 68, *them* 71, *my* 98, and (at .02) *how* 104. Whether the frequencies of all these words are the particular bias of M or of the Standard Corpus itself will appear when the frequency list of the Standard Corpus is published.

Reserving discussion of orders higher than the 122nd for the next section, we turn now to note some of the peculiarities of the other corpora. As we list words we will give their rank order in the M column to make it easier for the reader to find the words.

A remarkable peculiarity of *Ulysses* is the relatively low frequency of several verbs. Half of them are also low in the children's writing: *be* 15, *has* 30, *should* 91, and we may add *must* 87 (but only at the .05 level in U) and *were* 37 (but only at .05 in R). Those low in U but unremarkable in R include *is* 7, *are* 19, *have* 23, *will* 29, and *would* 39. Presumably the relatively low frequency of *is* and *are* is due to the heavier use which the largely expository corpora make of the timeless present in generalizations. The fact that *was* 13 stands at virtually identical ranks in all the corpora seems to confirm this supposition. The same may be said of the low frequencies of *have* and *has* as compared to the nearly identical frequency of *had* 34.

On the other hand, fiction has more need for *said* 51 than exposition has. Hence the rank orders of 72 and 75 in R and D, which are predominantly expository, and the order 23 in U. The fact that M is midway between the extremes suggests once more that it contains a disproportionately large share of the fiction of the Standard Corpus. We must come back to Joyce's verb system in the next section and will postpone further comment.

The fact that the children use *have* 23 and *had* 34 more frequently than the other corpora suggests that they need it in the sense of "possess" more often than adults. The reason why *has* 30 is com-

paratively low in frequency in their vocabulary is that they are more concerned with what "I and you" possess than with what "he, she, and it" do. This can be seen in the high frequency of the possessive determiners *my* 98 and *your* 68 and the low frequency of *his* 21 and *its* 48. While *her* 61 is at about the same rank level as in the adult writing, *she* 73 stands higher on the children's list than on the adult; this means that the children need *her* as the object pronoun oftener than the adults do and hence use it less often as a determiner. Both *their* 38 and *our* 65 are also low in frequency in R as compared to *they* 32 and *we* 35.

In R the orders of *is* 7, *was* 13, *are* 19, *will* 29, *can* 42, and *do* 72 are close to those of D and M. The orders of *be* 15, *were* 37, *been* 43, *may* 55, *could* 80, *should* 91, and *being* 118, and (at .05) *must* 87 are lower in rank than in M and D. The lower frequencies of *be* and *been* are partly a reflection of the lower frequencies of the modals and partly, along with the lower frequency of *being*, a reflection of the lesser use made by children of the more complex verb phrases. Children may not need the moral imperatives of *should* and *must* as often as adults, or they may substitute *have to* and *has to*. The logical use of *must* ("and it must follow as the night the day") and the probabilistic senses of *may* will appear less often in the immature style of children. If we add the instances of *couldn't* in the Rinsland corpus to the instances of *could*, the total would stand at the 84th rank, which is just between D and M. If we combine the same two words in D, they stand at the 82nd rank. The difference between the adult and children's use of the words is merely the greater frequency of the informal *couldn't* in the children's writing.

There are other words than verbs for which U and R share a difference from M and D. The lower frequency of *its* 48 reflects the more personal world of discourse of the novel and of the children. This is also the reason for the higher frequency of *she* 73. Though the world of children may be slightly less masculine than the others, the world of U is not. *he* 16, *him* 59, *you* 33 (also high in D and higher in R), *her* 61, *your* 68, and *I* 93 are all of comparatively high frequency in U; only *we* 35 and *our* 65 are comparatively low.

The greater concreteness of the children's and the novelist's worlds are suggested by the higher frequencies of the adverbs or prepositions of location and time: *out* 60, *up* 64, and *then* 79 (and for U alone we may add *over* 74, *after* 81, and *now* 82). We can discern differences of rhetoric in the lower frequencies of *for* 9, *which* 27, *since* 120, *own* 119 (but only at .01 in U), and for U alone, *but* 26, these words are not needed as much in narrative as in exposition, and they require a somewhat more complex style than all the children have mastered. (Of similar character is U's preference of *too* 117 over *also* 75.) R and U agree in giving *like* 103 a higher rank order. As a verb *like* would appear more frequently in personal than impersonal discourse; as a preposition of comparison it would be preferred to *such as* by both Joyce and the children. In her sleepward soliloquy, Molly Bloom's preposition is *like*, never *such as*. The lower frequency of *any* 77 suggests a lower frequency of generalizations in R and U.

The personal character of the children's world of discourse is also reflected in the strikingly low rank of *of* which is tenth in R and second in the other corpora. In a world in which personal nouns and proper names are specially frequent, the possessive inflection is more frequent and the periphrastic genitive with *of* less so.

Other prepositions of comparatively low rank in R are *with* 11, *by* 17, and *from* 22. The discrepancy is greatest for *by*, the agentive preposition in passive sentences, which do not occur so frequently in children's as in adult writing.

Of course *by* is also an adverb, but as such it has fewer meanings and enters into fewer idioms than do *to* 4, *in* 6, and *on* 14, which are in about the same rank in R as in the other corpora. Evidently in the syntax of children's writing even the simple prepositional phrase is less frequent than in adult writing.

In R *an* 28 is much lower in rank than elsewhere. The explanation seems to us to be partly the regional distribution of *a* before vowels and partly the age at which children master the alternation of *a* and *an* (if they ever do). The use of *a* before vowels is most common in the South Midland and South, where it is a nonstandard and receding feature, and in large urban centers. Since Rinsland was careful to draw upon all parts of the country and all kinds of schools for his corpus, these areas are well represented in R. Disadvantaged youths of college age sometimes substitute *a* for *an* in writing. We have also noticed that middle class children of well-educated parents have not fully mastered the use of *an* by age eight or nine, perhaps because it is the lone survivor of what was once a more extensive pattern in which *mine*, *thine*, and *none* also occurred before vowels and *my*, *thy*, and *no* before consonants.

We will mention that R substantiates the complaints of teachers that pupils overuse *very* 111 and begin too many sentences with *there* 45. But other points that might be made in explanation of the rank orders in R and U we will have to leave to the reader to make for himself.

III

Another way of establishing the bias of a corpus is to compare its most frequent words with those of other corpora. Of the first 122 in M, 76 words are common to the first 122 of all four corpora. (These 76 can be derived from Table 1 by noting which words have no number higher than 122 in the columns R, U, and D.) How the four lists of 122 words differ can be seen in Table 2, which lists the words uniquely present or absent in each corpus. (That is, under +R are listed the 24 words that appear in the 122 most frequent in R but not in the first 122 of any of the other three lists. Following the word is its rank order in R. Under -R appear the words that are in the first 122 of U, D, and M but not in the first 122 of R.)

TABLE 2

+R

school 26, am 29, got 34, went 36, dear 44, going 47, mother 64, friend 66, home 74, Christmas 79, write 81, play 83, came 84, put 85, house 87, saw 91, dog 102, name 107, want 110, soon 117, take 118, letter 120, sure 121, boy 122

—R

which 134, more 136, who 141, way 154, before 158, no 163, only 195, must 231, under 293, Mr. 301, its 312, those 491

+U

Bloom 30, Stephen 56, says 61, off 76, yes 77, eyes 80, O 81, hand 90, street 93, again 109*, face 111, right 113, round 115, head 119*

—U

been 126, can 128, made 145, make 157, much 162*, many 323, people 354

+D

war 55, men 72, great 85, upon 89, every 90, shall 94

—D

just 129, too 183

+M

also 75, last 83, years 84, Mrs. 92, each 96, year 99, state 100, work 114, use 115, three 116, own 119, since 120, still 121, used 122

—M

(R rank first, U second, D third)

come 67, 97, 92; did 75, 71, 118; down 78, 62, 120; here 94, 94, 105; I 2, 10, 10; know 109, 82, 115; little 54, 92, 99; me 27, 29, 47; see 65, 67, 113; us 88, 102*, 93

Other results of the comparison are presented in Table 3, which arrays the words common to any pair of lists and absent in the other pair. (Under +DM are the words that appear in the 122 most common in the Dewey and Monroe corpora but not in the 122 most common in *Ulysses* and the Rinsland corpus.) The three tables furnish the reader all the information he needs to derive and order the three lists of 122 words other than M, which is already given in Table 1.

TABLE 3

+DM

(D rank first, M second)

new 120, 44; than 66, 49; may 76, 55; these 81, 58; most 109, 70; such 88, 76; any 57, 77; even 122, 86; world 115, 88; should 96, 91; being 103, 118

+UM

(U rank first, M second)

through 108, 95; where 89, 107

+RU

(R rank first, U second)

go 41, 114; get 50, 101; back 98, 75; night 106, 112; old 108, 58

+UD

(U rank first, D second)

say 100, 100; never 104, 117; long 116, 122

+RD, +RM

No such sets.

To take the most obvious example first, it is easy to discern the bias of R and to account for it. The 140 or so adults who wrote M discussed so many different topics that, as we analyze in Table 4 the 122 most frequent words, only 20 of them are content words. The 100,000 and more children who wrote the Rinsland corpus lived in a much smaller world of discourse.

The words under +R of Table 2 describe this world: it has a school, a mother, a friend, a house, a home, Christmas, a dog, a name, a letter beginning with the word *dear*, and a boy. Its inhabitants write, play, put, want, take; they got, they came, they went, they saw . . . soon and sure.

The structure word *am* appears for use with *I*, which in the uninhibited writing of children is the word of second highest frequency (see last section of Table 2). In -R appear the title *Mr.* and the ten or eleven (if we count *way*) structure words crowded off the list by the content words of +R. It is interesting to note that children do not need *more* and *most* (+DM, Table 3) as often as adults, which suggests that they make heavier use of inflected adjectives. It is also interesting that they share with U a greater need for *old* (+RU) and a lesser need for *new* (+DM). On the whole the excluded words in -R and +DM suggest the more oral and simpler syntax of children's writing.

In the list under +U, *Bloom* and *Stephen* are obviously a bias of the novel. The Dublin mise en scène contributes *street*. *Eyes, hand, face, head* come partly from the particularity of fiction, partly from the highly personal nature of Joyce's narrative. The word *says* occurs 470 times in *Ulysses*, but at least 370 instances are in a 52-page chapter beginning, "I was just passing the time of day with old Troy" (Random House ed., pp. 287-339), where nearly every paragraph of conversation has the word *says*. Its appearance in +U is the result of the stylistic bias of this chapter. Also stylistic are *yes* and *O*. Molly Bloom is a highly affirmative woman: in the last chapter– her nocturnal soliloquy– she says yes at least 70 times and O at least 48.

Taking as his base the first hundred words in Dewey's frequency list, Martin Joos observed thirty years ago in Appendix II of Hanley's *Word Index* (p. 386) that Joyce was not very fond of the words *may*, *should*, *shall*, and *these*. From -U we can add *been*, *can*, *make* and *made*, and also *much* and *many*. From +DM we can add *being* to the other verbs, and from Table 1 we can add *this* 18 to *these*. If we add to the verbs in this paragraph the 9 others of comparatively low frequency cited in the previous section, it is apparent that the system of verb phrases in *Ulysses* differs markedly from that of the Dewey and Monroe corpora. How much of this to attribute to the differences in genre, how much to the difference between Irish English and American English, how much to the style of Joyce, and how much to the style of *Ulysses* are interesting questions that only another computer program could answer.

We now go on to +D. Dewey himself noted that *war* was unusually frequent because his corpus was collected in the spring of 1918, about half of it from newspapers and magazines. This also

accounts for the presence of *great* (as in the phrases "great war, great powers, great losses") and for *men* (as in discussions of military actions and losses). Some features of the corpus would exaggerate the frequency of *shall*: 5000 words of editorials from the Boston Evening Transcript, 5000 from Abraham Lincoln's speeches, 2000 from the sermons of Henry Ward Beecher and Phillips Brooks, etc. In our opinion, an additional reason for its absence from the other lists is a diminishment in its use in American English since 1918. Whether the presence of *upon* at rank 89 and the determiner *every* at rank 90 is due to a diachronic shift in the language or to the bias of the Dewey corpus we cannot say. We may know more about this when the frequency list for the whole Standard Corpus is published.

The absence of *too* (-D) is surely a result of bias in the corpus. But *just* at order 129 is almost on the list: if there had been more colloquial matter in the corpus, *just* would have been included.

As to the special bias of M, we note that there are 14 words in +M and only 6 in +D. Similarly there are 10 words in -M and only 2 in -D. Judged by this criterion as well as by the number of words that are of comparatively high frequency in M alone, M would appear to be further from the baseline of American English in respect to high-frequency words than D.

Having shown that the bias of a corpus is reflected even in its 122 most common words (though not much in the first 35 unless the corpus is quite peculiar) and having shown that some aspects of the bias may be detected by comparison with other frequency lists, we will proceed to hazard a few guesses about the 122 most frequent words of the Standard Corpus. We surmise that when the list is published the words *come*, *I*, *little*, *me*, and *us* will appear, and that the words *Mrs.*, *state*, *year*, *use* and *used* will not appear. As to the more disputable territory of *each*, *last*, *years*, and *own* as against *did*, *down*, *here*, *know*, and *see*, we think it likely that more of the last five will appear than of the first four.

IV

In Table 4 we array the 122 words of M in a grammatical classification to determine what per cent of the corpus is taken up by the several subclasses. To keep our groups from getting too small and too numerous we have had to make compromises and approximations. Other grammarians would have made other decisions: anyone who does not like our grouping can make his own from Table 1.

TABLE 4

Per cent of corpus			
10.7	A.	9	determiners: the, a, an, their, its, no, our, your, my
3.2	B.	11	determiners/pronominals: that, this, his, all, some, these, her, any, those, each, much
6.6	C.	7	prepositions: of, for, with at, from, into, like
6.8	D.	10	prepositions/adverbials: to, in, on, by, out, about, up, over, through, under
1	E.	9	adverbials: only, also, then, now, even, just, very, too, still
2.7	F.	8	personal pronouns: it, he, they, you, we, him, them, she
3.5	G.	3	correlatives: and, or, but
1	H.	5	relatives/interrogatives: which, who, what, how, where
1.7	I.	11	numerative adjectivals: one, more, other, first, two, most, such, many, last, three, own
0.8	J.	6	subordinators: if, when, after, because, before, since
4.7	K.	11	auxiliaries: is, was, be, are, have, has, had, were, been, do, being
1.3	L.	7	modals: will, would, can, may, could, must, should
1.7	M.	5	miscellaneous: as, not, there, than, so
45.8		102	structure words
2.1		20	content words: new, said, time, years, made, world, good, man, Mrs., Mr., year, state, people, way, make, well, day, work, use, used

We define determiners strictly so that there is only one prenominal determiner

to a noun phrase (including the zero allomorph of *a/an*); that is, determiners displace each other. Then the numerative adjectivals are the words that can occupy the slot between the determiner and the first veritable descriptive adjective--for example, *two* and *other* in the phrase "the two other friendly fellows." Except for the cardinal and ordinal numerals, which are theoretically infinite, this is a small, closed class. A few of our other groups are more of a mixture, some of their members having more than one or two grammatical functions.

The line between content words and structure words is admittedly a blurry one. The auxiliary *have* belongs on one side, the lexical *have* meaning "possess, etc." belongs on the other. We have assumed that the uses as an auxiliary in M outnumber the others. For *way*, on the other hand, we assumed that its content uses outnumbered its structural ones, as we did also for *used*. Our use of the term "content" in this sense is not intended to imply that structure words do not have meaning or content.

As we classify them, the determiners, the determiners that are also pronominals, the prepositions, and the prepositions that are also adverbials account for over a quarter of the whole corpus. Add to these the modals and auxiliaries, and 55 structure words account for exactly a third of the corpus. The 20 content words appear 5862 times, the 102 structure words 130,505 times. Since the 102 are less than a fifth of the structural vocabulary of English, it appears that more than half the words in a typical composite corpus will be structural.

The last lines of Table 4 call to our attention the fact that *time* is the noun of highest frequency in M, as was *Zeit* in a very large German corpus counted by F. W. Kaeding at the end of the last century.[7] We take it this signals common elements in western culture rather than the Germanic brotherhood of the languages. After *war*, *time* is the most frequent noun in D. In U, after *Bloom* and *Stephen* the first common noun is *man*, followed by *time*. In R the first is *school*, followed by *time*, and then by *mother*. Evidently the bias of a corpus can sometimes be discovered in its first noun alone.

One further point of interest: of the 122 words, only 12 are not of OE origin. *They*, *their*, *them* are early ME borrowings from ON. *Mr.* and *because* are respectively fusion and compound of OE and OF or Latin elements; and *Mrs.*, *state*, *people*, *just*, *very*, *use*, and *used* are of OF or Latin origin. All but *Mr.* (15th c.) and *Mrs.* (16th) are ME borrowings. All told, they occur 4172 times in the corpus (1.5 per cent of it) as against 132,195 occurrences of the native words (46.4 per cent of the corpus), which occur almost 32 times as often as the borrowed ones. Not only do structure words bulk large in a typical corpus; they are almost exclusively native English words.

[7] Ernest Horn. *A Basic Writing Vocabulary*. University of Iowa Monographs in Education. 1926, p. 189.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.